Automatic Image Colorization using Convolutional Neural Networks

Alexandre Dalban
CentraleSupélec
alexandre.dalban@student-cs.fr

Wacil Lakbir
CentraleSupélec
wacil.lakbir@student-cs.fr

December 17, 2025

Abstract

Automatic image colorization presents significant challenges due to the ambiguity of mapping grayscale intensities to plausible color distributions. This project investigates CNN-based approaches to colorization, comparing regression methods across different architectures (baseline CNN, U-Net) and color spaces (RGB, LAB), as well as exploring a classification-based formulation using discretized color bins. We train and evaluate our models on 5,800 high-quality images, analyzing how architectural choices, color space representations, and loss functions affect colorization quality. Our experiments show that U-Net architectures with skip connections provide notable improvements over baseline encoder-decoders, and that classification approaches can help address the color averaging problem inherent in regression methods, though trade-offs exist. Code: `https://github.com/ADnocap/Recolour-Greyscale.git`.

1 Introduction

Image colorization—the task of predicting plausible colors from grayscale images—has practical applications in historical photo restoration, artistic enhancement, and accessibility tools. While humans can infer colors from semantic understanding (grass is typically green, sky is typically blue), this task is challenging for machine learning systems due to its ill-posed nature: a single grayscale intensity can correspond to many different valid colors.

Convolutional Neural Networks have shown promise for colorization by learning semantic color associations from large datasets [10, 3]. However, a fundamental issue arises when using standard regression losses: mean squared error (MSE) tends to produce desaturated, brownish outputs when multiple plausible colors exist for the same grayscale region. This "color averaging" problem occurs because MSE optimization converges toward the conditional mean, effectively blending distinct color modes into muddy averages.

In this project, we conduct a systematic exploration of CNN-based colorization approaches. We investigate three key aspects: first, we compare architectural choices by testing a simple encoder-decoder CNN against U-Net architectures that incorporate skip connections for multi-scale feature preservation. Second, we examine the effect of color space representation by evaluating performance in RGB versus LAB color space, where LAB separates luminance from chrominance channels. Third, we analyze how different loss functions (MSE, L1, and their combinations) affect output quality and color saturation. Additionally, we implement and evaluate a classification-based approach that discretizes the color space into bins, aiming to better handle the multi-modal nature of the colorization problem.

2 Related Work

Deep learning for colorization. Early deep learning approaches to colorization employed feed-forward CNNs with regression losses [1]. Zhang et al. [10] introduced the idea of reframing colorization as classification over quantized color bins, which helped the model capture multi-modal color distributions. Iizuka et al. [3] incorporated both global and local features through a two-stream architecture. More recent work has explored generative adversarial networks [4, 8], self-attention mechanisms [7], and transformer architectures [5].

Loss functions. The choice of loss function has a significant impact on colorization quality. MSE loss is known to produce desaturated outputs by optimizing for the conditional mean [10], while L1 loss, which optimizes for the conditional median, can yield sharper colors but may introduce artifacts. Perceptual losses [6] and adversarial losses [2] have been used to improve visual quality, though they add training complexity.

Color spaces. While RGB is the standard color representation, LAB color space offers potential advantages by separating luminance (L) from chrominance (AB). This decom-

position allows models to focus specifically on color prediction while preserving the original brightness information [10, 1].

3 Method

3.1 Problem Formulation

Given a grayscale image $I_L \in \mathbb{R}^{H\times W}$, our goal is to predict a color image $I_{RGB} \in \mathbb{R}^{H\times W\times 3}$ or, when working in LAB space, the chrominance channels $I_{AB} \in \mathbb{R}^{H\times W\times 2}$. We frame this as a supervised learning problem, training CNNs on paired grayscale-color image datasets to learn the mapping from luminance to color.

3.2 Architectures

Baseline CNN. Our baseline architecture follows a simple encoder-decoder design with 4.3M parameters. The encoder consists of four downsampling blocks, each containing a convolutional layer followed by batch normalization, ReLU activation, and max pooling. This progressively reduces the spatial resolution from 96×96 to 6×6 while increasing the channel depth from 1 to 512 (1→64→128→256→512). The decoder mirrors this structure using transposed convolutions to upsample back to 96×96 resolution. A final sigmoid activation produces normalized RGB values in the range [0,1].

U-Net. To address the loss of spatial detail inherent in the baseline's bottleneck architecture, we adopt the U-Net design [9], which adds skip connections between encoder and decoder layers at matching resolutions. These skip connections are implemented via channel-wise concatenation, which doubles the channel count in decoder layers. This architecture preserves fine-grained spatial information that would otherwise be lost during downsampling, potentially enabling sharper colorization boundaries and better preservation of texture details. We implement two U-Net variants: a smaller RGB version with 4.5M parameters operating at 96×96 resolution with a 4-level encoder-decoder structure, and a larger LAB version with 20.5M parameters designed for high-resolution 512×512 colorization with a deeper architecture.

3.3 Color Space Representations

RGB approach. In the RGB formulation, the model receives a single-channel grayscale image as input and predicts all three RGB color channels. This means the network must jointly learn both brightness and color information, which may increase the difficulty of the learning task since the grayscale input already encodes brightness.

LAB approach. The LAB color space separates luminance (L channel, ranging 0-100) from chrominance (A and B channels, ranging approximately -128 to +127, representing green-red and blue-yellow color axes respectively). We extract the L channel from RGB images to use as input and train the model to predict only the AB channels. At inference time, we recombine the predicted AB channels with the original L channel to reconstruct the full RGB image. This decoupling simplifies the learning objective by removing the need to predict brightness information. The complete pipeline involves: (1) loading RGB images at 512×512 resolution, (2) converting to LAB color space, (3) normalizing the L channel to [0,1] for model input, (4) normalizing AB channels to [-1,1] as training targets, (5) predicting AB values through the network, and (6) reconstructing RGB by combining predicted AB with input L.

3.4 Loss Functions

The averaging problem. To understand why standard regression losses can be problematic for colorization, consider a grayscale pixel with intensity $I_L = 128$ that corresponds to three different colors in the training set: red (255,0,0), green (0,255,0), and blue (0,0,255). When using MSE loss:

$$\mathcal{L}_{MSE} = \mathbb{E}_{(x,y)}[\|f(x) - y\|^2] \tag{1}$$

the model's optimal prediction under this loss is the conditional mean:

$$f^*(128) = \mathbb{E}[RGB|I_L = 128] = (85, 85, 85) \tag{2}$$

This results in a muddy brown color that averages the three pure hues, rather than committing to one of the plausible options.

L1 loss. An alternative is L1 loss, which optimizes for the conditional median:

$$\mathcal{L}_{L1} = \mathbb{E}_{(x,y)}[\|f(x) - y\|_1] \tag{3}$$

In the RGB example above, L1 would predict the per-channel median of (0,0,0). While this doesn't fundamentally solve the multi-modality issue, L1 loss empirically tends to produce somewhat sharper and more saturated colors compared to MSE.

Combined loss. To balance the smoothness properties of MSE with the sharpness-inducing properties of L1, we use a weighted combination:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{MSE} + (1 - \alpha)\mathcal{L}_{L1} \tag{4}$$

We experiment with different weighting factors $\alpha \in \{0.3, 0.5, 0.7\}$ across our models to find a good balance.

3.5 Classification-Based Colorization

To explore an alternative approach that may better handle multi-modal color distributions, we implement a classification-based method inspired by Zhang et al. [10]. Instead of regressing continuous AB values, we discretize the AB color space into bins and treat colorization as a classification task.

Color space quantization. The AB color space spans [-128, 127] in both dimensions. We partition this continuous space into a grid of discrete bins with bin size $b = 10$, yielding 26×26=676 total bins. Each bin center represents a quantized color:

$$\text{bin\_centers} = \{(a, b) : a, b \in \{-125, -115, ..., 125\}\} \tag{5}$$

During training, each pixel's ground truth AB color is assigned to its nearest bin using Euclidean distance, creating a hard classification target.

Architecture modification. We use the same U-Net encoder-decoder structure as our LAB regression model, but modify the output layer to produce a probability distribution over all possible color bins:

$$f : \mathbb{R}^{H \times W} \rightarrow \mathbb{R}^{H \times W \times 676} \tag{6}$$

The network predicts a probability distribution over the 676 bins for each pixel via a softmax activation.

Loss function. We employ weighted cross-entropy loss to account for class imbalance, as some colors appear much more frequently in natural images than others:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N} w_{y_i} \log p(y_i|x_i) \tag{7}$$

where w_c are empirical class weights computed from training data frequency to give more importance to underrepresented colors.

Inference. At test time, we take the argmax over the predicted bin probabilities and map back to continuous AB values using the bin centers:

$$\hat{AB}_{i,j} = \text{bin\_center}[\arg\max_k p(k|L_{i,j})] \tag{8}$$

This formulation theoretically allows the network to represent multi-modal distributions: if red, green, and blue are all plausible for a given grayscale value, the network can assign probability mass to multiple bins without being forced to produce a single averaged prediction.

4 Experiments

4.1 Dataset

We constructed a custom dataset using the Pexels API, selecting high-quality, color-rich images suitable for training colorization models. The dataset contains 5,800 training images and 1,000 test images at 512×512 resolution, covering diverse categories including people, places, natural scenes, and objects. Images were manually curated to ensure color diversity and visual quality. For the 96×96 models, we perform downscaling during data loading. Grayscale versions are generated by extracting the luminance channel (either L from LAB or weighted RGB-to-grayscale conversion).

4.2 Training Setup

For the Baseline CNN and U-Net RGB models operating at 96×96 resolution, we use 4.3M and 4.5M parameters respectively, training with the Adam optimizer at learning rate 1e-4 and a batch size of 512 (enabled by the smaller resolution). Both are trained for 20 epochs with a combined loss of 0.7 MSE + 0.3 L1. The larger batch size provides more stable gradient estimates during optimization.

The U-Net LAB model operates at full 512×512 resolution with 20.5M parameters. Due to memory constraints at this higher resolution, we reduce the batch size to 16 while maintaining the Adam optimizer at learning rate 1e-4. This model trains for 15 epochs with a balanced loss weighting of 0.5 MSE + 0.5 L1, which we found empirically to work well for the AB color space.

For the classification-based U-Net, we use the same 20.5M parameter architecture as the LAB model but replace the regression head with a 676-class classification head (one class per color bin). Training uses Adam with learning rate 1e-4, batch size 16, and weighted cross-entropy loss with class weights computed from the training set color distribution. We train this model for 20 epochs.

4.3 Evaluation Metrics

Peak Signal-to-Noise Ratio (PSNR). We use PSNR as our primary quantitative metric, which measures reconstruction quality in decibels:

$$PSNR = 10 \log_{10} \left(\frac{MAX^2}{MSE} \right) \tag{9}$$

where $MAX = 1.0$ for normalized images. Higher PSNR values indicate better pixel-wise accuracy. Typical interpretation ranges are: less than 20 dB indicates poor quality, 20-30 dB is acceptable, and above 30 dB is considered good quality.

However, PSNR has important limitations. While it correlates with MSE and is widely used for its simplicity, it doesn't always reflect perceptual quality well. Two colorizations with identical PSNR may differ significantly in visual appeal, particularly regarding color saturation and semantic correctness. We therefore supplement PSNR with qualitative visual analysis.

5 Results

5.1 Quantitative Results

Table 1 presents the quantitative performance of our models on the 1,000-image test set. We observe that the U-Net architectures consistently outperform the baseline CNN, with U-Net RGB achieving 13.5 dB PSNR compared to 11.2 dB for the baseline—a gain of 2.3 dB. The U-Net LAB model, operating at higher resolution, reaches 19.0 dB. The classification-based approach achieves 18.2 dB mean

Table 1: **Quantitative results** on the test set (1,000 images). MSE is computed on normalized [0,1] images for regression models. Note that resolution and color space differ across models.

Model	Res.	PSNR (dB) ↑	Loss
CNN Baseline	96×96	11.2	Regression
U-Net RGB	96×96	13.5	Regression
U-Net LAB	512×512	19.0	Regression
U-Net Classif.	512×512	18.2	Classification

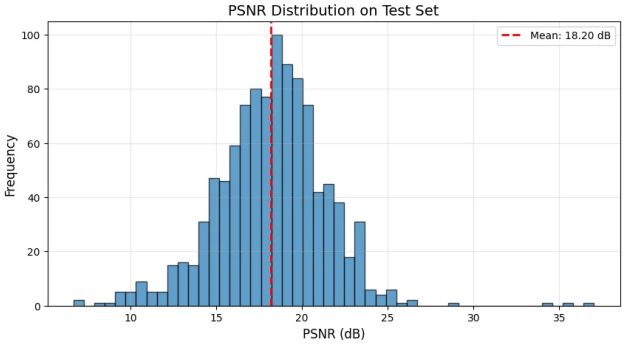


Figure 1: **PSNR distribution for classification model.** The histogram shows PSNR values across 1,000 test images with a mean of 18.20 dB. Most images concentrate between 17-21 dB with few extreme outliers.

PSNR, which is comparable to the LAB regression model's pixel-wise accuracy.

Figure 1 shows the PSNR distribution for the classification model across the test set. The distribution is reasonably concentrated with most images falling between 17-21 dB and relatively few outliers, suggesting consistent performance across different image types.

5.2 Qualitative Analysis

Figure 4 presents visual comparisons of colorizations from all four models. The baseline CNN produces heavily desaturated, brownish results with noticeable color bleeding across object boundaries. Fine spatial details are lost due to the severe bottleneck at 6×6 resolution. The U-Net RGB model shows clear improvement in both color saturation and spatial localization—the skip connections help preserve edge information and reduce color bleeding across boundaries. However, some brownish averaging still appears in ambiguous regions.

The U-Net LAB model produces more vibrant colors and sharper boundaries at the higher 512×512 resolution. The LAB formulation appears to help the model focus its capacity on chrominance prediction. That said, the model sometimes over-saturates certain regions and still exhibits averaging behavior in scenarios with multiple plausible colors.

The classification-based U-Net produces noticeably more saturated and confident colors compared to the regression models. By representing color as a distribution over discrete bins, the model can commit to specific color choices rather than averaging. Colors appear more natural and semantically plausible, particularly for objects with strong color priors like grass, sky, and skin tones. However, careful examination reveals some quantization artifacts as discrete boundaries between color bins, especially visible in smooth gradient regions.

5.3 Ablation Studies

To isolate the effect of skip connections, we compare the CNN Baseline (no skip connections) to U-Net RGB (with skip connections) at identical 96×96 resolution. The 2.3 dB PSNR improvement demonstrates that preserving spatial information through skip connections provides meaningful benefits for colorization quality.

Regarding color space, the LAB formulation reduces problem complexity by separating luminance from chrominance. While direct comparison is complicated by resolution differences, qualitative results suggest that LAB may enable more saturated colorizations by focusing the model's capacity specifically on color prediction.

Comparing regression versus classification, we find that U-Net LAB (regression) achieves 19.0 dB PSNR while U-Net Classification achieves 18.2 dB. The similar pixel-wise accuracy but visually different results highlight PSNR's limitations—it measures pixel-level error but doesn't capture perceptual qualities like color saturation or semantic appropriateness. The classification model produces more confident, saturated colors by avoiding the conditional mean convergence that regression losses exhibit.

For loss weighting in regression models, we trained U-Net RGB with different α values (MSE weight in combined loss). Higher MSE weight ($\alpha = 0.7$) produces smoother but more desaturated colors, while lower weight ($\alpha = 0.3$) increases saturation but can introduce artifacts. The balanced $\alpha = 0.5$ provides a reasonable trade-off.

For the classification approach, we experimented with different bin sizes $b \in \{5, 10, 15\}$. Smaller bins ($b = 5$, yielding 2704 classes) provide finer color granularity but substantially increase training difficulty and memory requirements. Larger bins ($b = 15$, yielding 196 classes) are easier to train but produce more visible quantization artifacts. The bin size $b = 10$ (676 classes) represents a practical balance between granularity and tractability.

5.4 Limitations and Failure Cases

Our approaches face several limitations. First, the regression models continue to produce desaturated averages when multiple distinct colors share similar grayscale values, de-

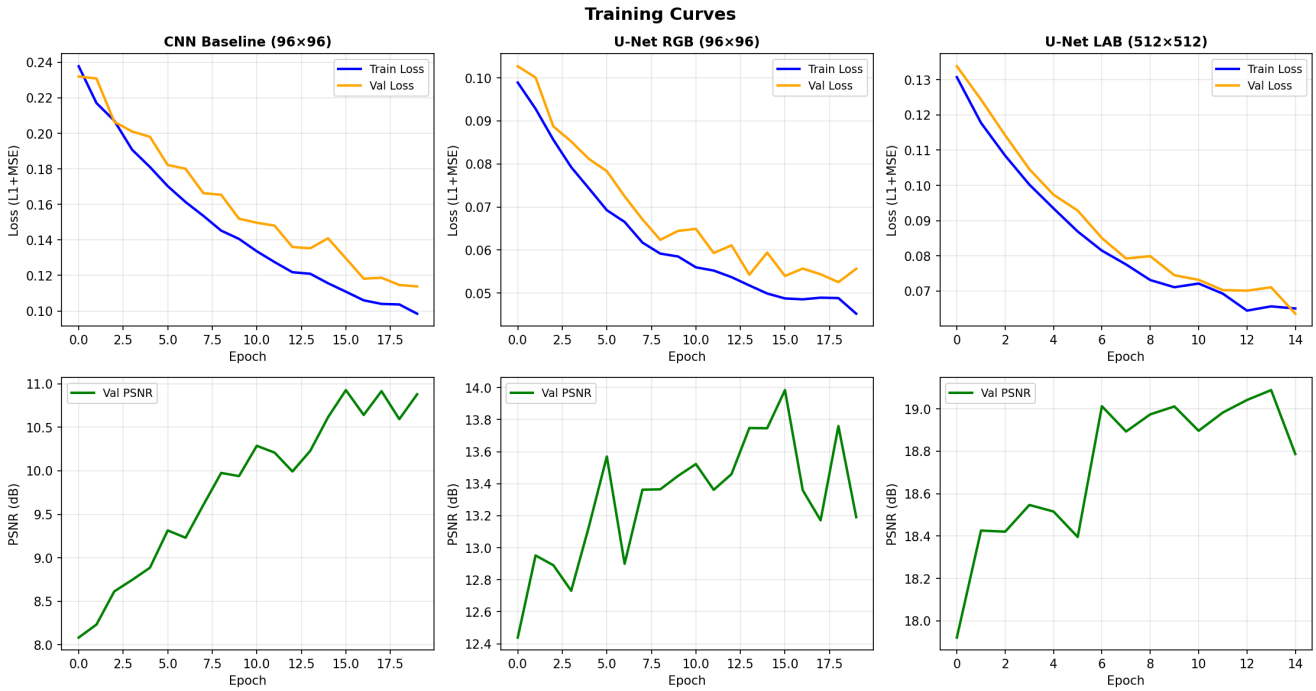


Figure 2: **Training curves for regression models.** Loss and PSNR evolution during training for (a) CNN Baseline, (b) U-Net RGB, and (c) U-Net LAB. The U-Net models show faster initial convergence and somewhat more stable validation metrics compared to the baseline. The U-Net LAB model achieves the highest PSNR among our regression approaches.

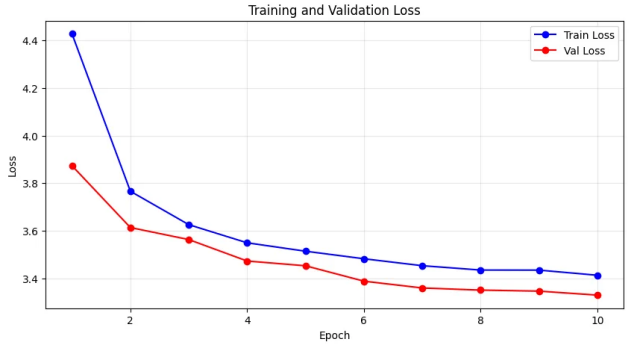


Figure 3: **Classification model training curves.** Training and validation loss (weighted cross-entropy) over 10 epochs. The curves show convergence with validation loss stabilizing around 3.35 after approximately 8 epochs, suggesting the weighted loss successfully handles class imbalance to some degree.

spite using combined L1+MSE losses. The classification approach helps mitigate this but introduces its own issues.

Second, the classification model sometimes produces visible quantization artifacts—discrete boundaries between color bins are particularly noticeable in smooth gradients like skies or skin. Finer bin sizes or post-processing techniques could potentially reduce these artifacts.

Third, all our models struggle with objects that have no strong color priors in natural images, such as painted walls, synthetic objects, or clothing items that come in many colors. Even classification cannot resolve ambiguity when semantic context is insufficient.

Fourth, the classification model has difficulty with rare colors due to class imbalance, despite our use of weighted loss. Very infrequent color bins (such as unusual purple or orange tones) remain underrepresented in training data and are predicted less reliably. These failure cases indicate that while our architectural and formulation choices provide improvements, fundamental challenges remain in the colorization problem that would require additional techniques to address.

6 Discussion

Our experiments confirm that architectural choices significantly impact colorization performance. U-Net architectures with skip connections provide substantial improvements over simple encoder-decoders by preserving spatial information across resolution scales. The LAB color space representation shows promise by decoupling luminance from chrominance, potentially allowing the model to focus its capacity more effectively.

The classification-based approach represents an interesting alternative formulation. Where regression with MSE loss optimizes for $\mathbb{E}[Y|X]$ and thus predicts the conditional mean, classification with cross-entropy can represent a full distribution $p(Y|X)$ over plausible colors. Taking the argmax at inference selects the most likely color without averaging, which can enable more confident, saturated predictions. However, this comes with trade-offs: quantization artifacts from discrete bins, increased difficulty handling class imbalance, and higher memory requirements.

The choice of bin size presents a clear trade-off between color granularity and training tractability. Future work could explore soft binning schemes, learnable color centers that adapt during training, or hierarchical classification strategies to better handle rare colors. Additionally, probabilistic decoding that samples from the predicted distribution rather than simply taking the argmax might produce more diverse colorizations.

Beyond the techniques explored in this project, several directions could further improve results. Perceptual losses based on features from pre-trained networks [6] might better capture semantic correctness. Interactive user guidance through sparse color hints [11] could help resolve ambiguities. Generative models like GANs or diffusion models might produce even more realistic colorizations, though at the cost of increased training complexity.

Our study has several limitations. The dataset size (5,800 images) is relatively modest compared to large-scale benchmarks used in recent literature. We did not extensively explore post-processing techniques that might reduce quantization artifacts in the classification approach. Finally, our reliance on PSNR as the primary quantitative metric is imperfect, as it doesn't always correlate well with perceptual quality—future work should incorporate perceptual metrics or user studies.

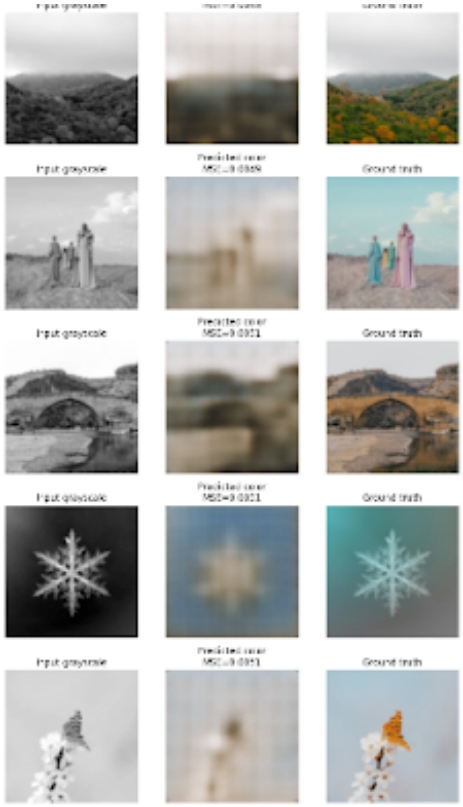
7 Conclusion

This project presented a systematic exploration of CNN-based automatic image colorization. We compared architectural choices (encoder-decoder versus U-Net), color space representations (RGB versus LAB), loss functions (MSE, L1, and combinations), and problem formulations (regression versus classification). Our experiments demonstrate that U-Net architectures provide meaningful improvements over baseline CNNs through multi-scale feature preservation, achieving approximately 2 dB PSNR gains. We provided detailed analysis of the color averaging problem inherent to regression losses and showed that combined L1+MSE losses offer a practical compromise between saturation and stability.

We also explored reformulating colorization as classification over discretized color bins. This approach showed improvements in color saturation and confidence compared to regression baselines by representing probability distributions over plausible colors rather than committing to a single averaged prediction. However, challenges remain including quantization artifacts and difficulty handling rare colors due to class imbalance.

Overall, this project illustrates the multiple design decisions involved in building colorization systems and highlights the trade-offs between different approaches. While we made progress in addressing some challenges, automatic colorization remains a difficult problem with room for further improvement through more sophisticated architectures, larger datasets, and alternative training strategies.

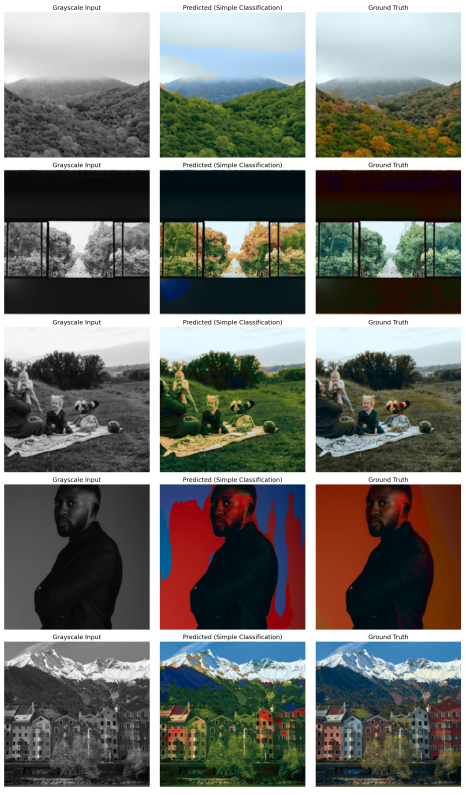
Appendix



(a) CNN Baseline

(b) U-Net RGB

(c) U-Net LAB



(d) U-Net Classification

Figure 4: Qualitative comparison of the four colorization models.

A Selecting the mixing weight α for the combined loss

A.1 Link to Section 3.4: mitigating the averaging problem

As discussed in Section 3.4, colorization is inherently *multi-modal*: the same grayscale input may correspond to multiple plausible colors. Under the MSE (L2) objective, the model is encouraged to approximate the conditional mean, which can lead to *averaging* of plausible hues and thus to desaturated, "muddy" predictions. In contrast, L1 (absolute error) tends to favor sharper and more saturated outputs in practice, although it does not fully solve multi-modality.

To balance the smoothness of MSE with the sharpness-inducing behavior of L1, we use the combined loss:

$$\mathcal{L}_{\text{total}}(\alpha) = \alpha\,\mathcal{L}_{\text{MSE}} + (1-\alpha)\,\mathcal{L}_{\text{L1}}, \qquad \alpha \in [0,1]. \quad (10)$$

A.2 Protocol and metric

We experiment with $\alpha \in \{0.3,\ 0.5,\ 0.7\}$, and include the limiting cases $\alpha = 1$ (MSE-only) and $\alpha = 0$ (L1-only). All runs use the same architecture and training hyperparameters to ensure a fair comparison.

Since PSNR is derived from MSE, we compute PSNR from the validation/test MSE (independently of the training loss):

$$\text{PSNR} = 10\log_{10}\left(\frac{1}{\text{MSE}+\varepsilon}\right), \qquad (11)$$

assuming images are normalized to $[0,1]$.

A.3 Why we retain $\alpha = 0.5$ and $\alpha = 0.7$

Figure 5 reports the validation PSNR curves across epochs for the different training losses. It shows that all settings converge to a similar PSNR range, but the mixed losses with $\alpha = 0.5$ and $\alpha = 0.7$ remain consistently among the best-performing curves while still incorporating an L1 component, which directly addresses the averaging issue described in Section 3.4.

Consistently with the trend observed in Figure 5, our final run yields the following best-checkpoint test PSNR values:

- **MSE-only** ($\alpha = 1.0$): PSNR $\approx$ 22.89 dB (corresponding to MSE ≈ 0.00514),
- **L1-only** ($\alpha = 0$): PSNR $\approx$ 22.85 dB,
- **Mixed** ($\alpha = 0.5$): PSNR $\approx$ 22.94 dB,
- **Mixed** ($\alpha = 0.7$): PSNR $\approx$ 22.92 dB.

These results indicate that adding an L1 component does not degrade reconstruction fidelity measured by PSNR; on the contrary, the mixed objectives slightly improve PSNR compared to MSE-only in this run. At the same time, the presence of the L1 term helps reduce the tendency of pure MSE to average multiple plausible colors, which is precisely the failure mode highlighted in Section 3.4.

Therefore, we keep two operating points on the fidelity–sharpness trade-off:

- $\alpha = 0.5$ as our *main balanced setting* (best PSNR among the tested configurations while mitigating averaging),
- $\alpha = 0.7$ as a *more MSE-oriented alternative* that remains close in PSNR while still benefiting from the L1 component.

For the remainder of the experiments, we report results for $\alpha \in \{0.5,\ 0.7\}$.

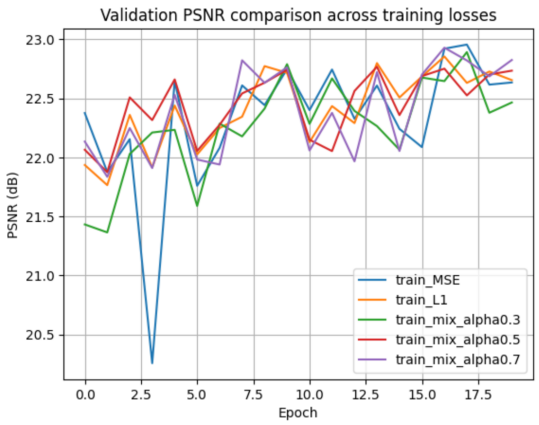


Figure 5: Validation PSNR across epochs for different training losses (MSE, L1, and mixed losses with $\alpha \in \{0.3, 0.5, 0.7\}$).

References

[1] Zezhou Cheng, Qingxiong Yang, and Bin Sheng. Deep colorization. *Proceedings of the IEEE International Conference on Computer Vision*, pages 415–423, 2015.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[3] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Let there be color! joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. In *ACM Transactions on Graphics (ToG)*, volume 35, pages 1–11. ACM, 2016.

[4] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[5] Xiaozhong Ji, Boyuan Jiang, Donghao Luo, Guangpin Tao, Wenqing Chu, Zhifeng Xie, Chengjie Wang, and Ying Tai. Colorformer: Image colorization via color memory assisted hybrid-attention transformer. In *European Conference on Computer Vision*, pages 20–36. Springer, 2022.

[6] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[7] Manoj Kumar, Dirk Weissenborn, and Nal Kalchbrenner. Colorization transformer. *International Conference on Learning Representations*, 2021.

[8] Kamyar Nazeri, Eric Ng, and Mehran Ebrahimi. Image colorization using generative adversarial networks. In *International conference on articulated motion and deformable objects*, pages 85–94. Springer, 2018.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[10] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. *European conference on computer vision*, pages 649–666, 2016.

[11] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S Lin, Tianhe Yu, and Alexei A Efros. Real-time user-guided image colorization with learned deep priors. In *ACM Transactions on Graphics (TOG)*, volume 36, pages 1–11. ACM, 2017.